IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 10, October 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

Applying Naive Bayes Techniques for Accurate Sentiment Analysis in Movie Reviews

Prof. Saurabh Sharma, Prof. Vishal Paranjape, Prof. Zohaib Hasan

Dept. of Computer Science & Applications, Baderia Global Institute of Engineering & Management, Jabalpur, India

ABSTRACT: This study examines the effectiveness of Naive Bayes and Logistic Regression classifiers in analyzing the sentiment of movie reviews. Two feature extraction approaches, namely Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), are utilized. We employed a dataset comprising 50,000 IMDB reviews that underwent preprocessing techniques such as denoising, stop word removal, and stemming. The reviews were transformed into vectors using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) approaches. Our investigation demonstrates that Logistic Regression surpasses Naive Bayes in terms of accuracy. Logistic Regression achieves 89.52% accuracy for Bag-of-Words (BoW) and 89.23% accuracy for Term Frequency-Inverse Document Frequency (TF-IDF), while Naive Bayes achieves 85.01% accuracy for BoW and 85.74% accuracy for TF-IDF. Naive Bayes has consistent performance with a minimum disparity between training and testing accuracies, indicating strong generalization skills despite its slightly lower accuracy. The results suggest that Logistic Regression outperforms Naive Bayes in terms of accuracy. However, Naive Bayes remains a strong contender because to its simplicity and consistent performance across various feature extraction methods. This comparison offers significant insights for choosing suitable classifiers and feature extraction techniques for text classification problems in sentiment analysis.

KEYWORDS: Sentiment Analysis, Naive Bayes, Logistic Regression, Bag of Words (BoW), TF-IDF, Text Classification, IMDB Reviews

I. INTRODUCTION

The most prevalent type of Naïve Bayes classifier is the multinomial naive Bayes classifier. It is named as such because it is a Bayesian classifier that simplifies the assumption about the interaction between features. Naive Bayes is a machine learning technique that utilizes Bayes' Theorem to make probabilistic predictions. It is commonly employed for classification purposes. Although it is simple, it is highly efficient in a wide range of applications including spam filtering, text categorization, sentiment analysis, and recommendation systems.

The approach is referred to as "naive" since it believes that the features utilized in the model are independent of each other, given the class label. Although this assumption is frequently impractical in real-life situations, it streamlines the calculation process and enhances the efficiency of the algorithm [1].

Naive Bayes is a probabilistic classifier that determines the class ĉ with the highest posterior probability for a given document d, out of all possible classes c ∈ C. In Equation 1, we utilize the hat notation ˆ to represent "our estimation of the accurate category".

Bayes' Theorem offers a method to revise the probability estimation for a hypothesis as further data or information is obtained. Mathematically, the expression is:

$$\hat{c} = argmax P(d) \quad (c \in C) \tag{1}$$

$$P(B) = \frac{P(A) \cdot P(A)}{P(B)} \tag{2}$$

In the context of Naive Bayes classification, (A) represents the class label, and (B) represents the features. The goal is to find the probability of a class given the features, which can be used to make predictions. There are several variants of the Naive Bayes classifier [2], including:

- **Multinomial Naive Bayes**: Commonly used for text classification, where the features are the frequency of words [3].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

- **Bernoulli Naive Bayes**: Suitable for binary/Boolean features [4].
- **Gaussian Naive Bayes**: Assumes that the features follow a normal distribution, often used for continuous data.

The algorithm works by calculating the posterior probability for each class and selecting the class with the highest probability. The steps involved include [3]:

- **Training Phase**: Calculate the prior probability for each class and the likelihood of each feature given the class.
- **Prediction Phase**: Use the calculated probabilities to predict the class label for new data.

Despite its "naive" assumption, Naive Bayes performs surprisingly well in many real-world applications, especially when the dimensionality of the data is high. Its simplicity, speed, and effectiveness make it a popular choice for initial classification tasks and baseline models.

Equation 2 can be substituted in equation 1 to get the following:

$$\hat{c} = argmax \frac{P(c) \cdot P(c)}{P(d)} \quad (c \in C) \tag{3}$$

P (d) can be dropped because P (d) doesn't change for each class; we are always asking about the most likely class forthe same document d, which must have the same probability P (d).

$$\hat{c} = argmax P(c) \cdot P(c) \quad (c \in C) \tag{3}$$

The selection of the Naive Bayes method is determined by its rapidity, simplicity of implementation, and efficacy, especially in datasets with numerous dimensions, owing to the assumption of feature independence [5]. In their study, the authors examined the performance of the Naive Bayes algorithm and the Support Vector Machine (SVM) on different datasets. They discovered that Naive Bayes frequently achieved superior accuracy compared to SVM. Similarly, a study was undertaken by [7] to compare the performance of Naive Bayes, k-nearest neighbor, and random forest algorithms in sentiment analysis of movie reviews. Their results demonstrated that Naive Bayes had superior performance in terms of accuracy compared to both k-nearest neighbor and random forest.

An inherent challenge in sentiment analysis is the abundance of characteristics, which can have a detrimental effect on the accuracy of categorization. In order to tackle this issue, it is crucial to implement a feature selection procedure. Chi-square is a widely used and successful feature selection tool among other methodologies [9].

The researchers conducted a study where they examined the Amazon Review Dataset, IMDb Review Dataset, and Yelp Review Dataset. They utilized various feature selection methods such as odds ratio, chi-square, GSS coefficient, and Bi-Normal Separation. These approaches were implemented using a range of algorithms, including logistic regression, SVM-RBF, SVM-Linear, decision tree, multinomial naïve Bayes, and Bernoulli naïve Bayes. Their research showed that the combination of multinomial naïve Bayes and chi-square feature selection provided the highest level of accuracy when applied to the Amazon Review Dataset and the IMDb Review Dataset.

II. LITERATURE REVIEW

The Naive Bayes classifier, originally proposed by Bayes in 1968, has played a fundamental role in the domains of text classification and sentiment analysis. The combination of its simplicity and effectiveness has resulted in its widespread adoption for a variety of applications. In their study, Abbas et al. (2019) investigated the application of the Multinomial Naive Bayes (MNB) model in sentiment analysis. They emphasized its effectiveness in dealing with extensive datasets and its reliability in categorizing textual data. Their study highlighted the efficacy of MNB in sentiment classification of movie reviews, establishing a robust basis for future research in this domain.

Kibriya et al. (2005) reexamined the use of MNB for text categorization, highlighting its effectiveness and precision. The researchers conducted a comparison between MNB and other classifiers and observed that MNB performed well, particularly in situations involving high-dimensional data. In their study, Singh et al. (2019) performed a comparative analysis of Multinomial and Bernoulli Naive Bayes classifiers for the purpose of text classification. Their research revealed that, overall, the Multinomial Naive Bayes (MNB) model performs better than the Bernoulli Naive Bayes model in terms of both accuracy and computational efficiency.

Taheri and Mammadov (2013) proposed optimization methods to improve the learning process of the Naive Bayes classifier. Their methodology sought to enhance the effectiveness of the classifier by refining its parameters, leading to a more precise sentiment classification. Jagdale et al. (2019) utilized machine learning methods, such as Naive Bayes,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

to do sentiment analysis on product reviews. Their research emphasized the significance of feature selection and preprocessing in attaining a high level of accuracy in classification.

Baik et al. (2017) conducted a study on sentiment analysis of movie reviews using different machine learning classifiers, including Naive Bayes. Their research emphasized the efficacy of Naive Bayes in managing varied datasets and its capacity to offer dependable sentiment predictions. In their study, Madasu and Elango (2020) examined effective methods for selecting features in sentiment analysis. They found that careful feature selection can greatly improve the performance of Naive Bayes classifiers.

III. METHODOLOGY

Naive Bayes Algorithm for Sentiment Classification

Problem Setup

1. Dataset: Let $\mathcal{D} - \{(x_i, y_i)\}_{i-1}^{N}$ be a set of movie reviews, where x_i is a feature vector representing the review and $y_i \in \{C_1, C_2, \dots, C_K\}$ is the sentiment label (class) for review i. We have K sentiment classes.
2. Feature Representation: Represent each review x_i as a vector of features $(x_{i1}, x_{i2}, \dots, x_{iM})$, where x_{ij} is the presence or frequency of the j-th feature (e.g. words or phrases).

Mathematical Formulation

1. Objective: Given a new review x, predict the sentiment class C that maximizes the posterior probability $P(C \mid x)$.
2. Bayes' Theorem: The posterior probability of a class C given features x is computed using Bayes' Theorem:

$$P(C \mid x) - \frac{P(x \mid C) \cdot P(C)}{P(x)}$$

where $P(x \mid C)$ is the likelinood of features given the class, $P(C)$ is the prior probability of the class, and $P(x)$ is the evidence.

3. Naive Assumption: The Naive Bayes classifier assumes that features are conditionally independent given the class C. Therefore,

$$P(x \mid C) - \prod_{j-1}^{M} P(x_j \mid C)$$

4. Likelihood Estimation: Estimate the likelihood $P(x_j \mid C)$ based on feature frequencies in the training data. For a discrete feature representation (e.g., word presence), the likelihood is:

$$P(x_j \mid C) - \frac{N_{jC} + \alpha}{N_C + \alpha \cdot M}$$

where N_{jC} is the count of feature j in documents of class C,N_C is the total count of features in documents of class C, α is a smoothing parameter (e.g., Laplace smoothing), and M is the total number of features.

5. Prior Probability: Estimate the prior probability $P(C)$ as the proportion of documents in class C:

$$P(C) - \frac{N_C}{N}$$

where N_C is the number of documents in class C, and N is the total number of documents.

6. Classification Rule: To classify a new review x, compute the posterior probability for each class C_k and select the class with the highest probability:

$$\hat{C} - \arg\max_{C_k} P(C_k \mid x) - \arg\max_{C_k} \left(\prod_{j-1}^{M} P(x_j \mid C_k) \cdot P(C_k) \right)$$

Detailed Algorithm Steps

1. Training Phase:

- Compute $P(C_k)$ for each class C_k.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

- For each feature x_j, compute $P(x_j \mid C_k)$ for each class C_k using the training data.

2. Prediction Phase:

- For a new review x, calculate the posterior probability for each class using the formula above.
- Choose the class with the highest posterior probability as the predicted sentiment.

Summary

This Naive Bayes algorithm for sentiment classification uses probabilistic principles to classify movie reviews into sentiment categories. It assumes feature independence given the class, which simplifies the computation and estimation of probabilities. This mathematical formulation provides a clear basis for implementing and understanding Naive Bayes methods in sentiment analysis.

The dataset used is IMDB Movie Reviews (TABLE 1)

TABLE 1 SAMPLE ROWS IN THE DATASET

Movie Review	**Sentiment**
You know that mouthwash commercial where the guy has a mouth full of Listerine or whatever it is and …	negative
I can't believe it that was the worst movie i have ever seen in my life. i laughed a couple of times …	negative
This is one of a rarity of movies, where instead of a bowl of popcorn one should watch it with a bot …	negative
Even though I'm quite young, The Beatles are my ABSOLUTELY FAVOURITE band! I never had the …	positive
An American Werewolf in London had some funny parts, but this one isn't so good. The computer …	negative
Originally I was a Tenacious D fan of their first album and naturally listened to a few tracks off …	positive
This is just one more of those hideous films that you find on Lifetime TV which portray the …	negative
The premise of this movie was decent enough, but with subpar acting, it was just bland and dull….	negative
The Lives of the Saints starts off with an atmospheric vision of London as a bustling city of busy …	negative
I can't emphasize it enough, do *NOT* get this movie for the kids. For that matter, …	negative

The process commenced by loading and examining the IMDB dataset, which included of movie reviews and their associated sentiments, with the pandas library. The initial data exploration process entailed analyzing the organization and dispersion of the dataset in order to acquire valuable insights. Subsequently, the data preprocessing phase commenced, encompassing multiple phases. The HTML tags were removed using the BeautifulSoup library, while the text enclosed in square brackets was eliminated using regular expressions. The text was further cleaned by removing special characters, and the Porter Stemmer from NLTK was used to reduce words to their base forms. In addition, frequently used English stopwords were eliminated in order to decrease the amount of irrelevant information in the text.

The dataset was partitioned into training and testing sets, with the initial 40,000 reviews allocated for training and the remaining reviews for testing. Two approaches, Bag of Words (BoW) and TF-IDF, were used for feature extraction. The CountVectorizer was employed to transform textual data into numerical characteristics using n-grams (specifically, unigrams, bigrams, and trigrams) for the Bag-of-Words (BoW) methodology. Additionally, the TfidfVectorizer was employed to implement the TF-IDF methodology, which takes into account the significance of words inside the documents.

The sentiment labels were subjected to label encoding, which transformed them into binary values of 0 (representing negative) and 1 (representing positive) using the LabelBinarizer. Three distinct models were trained utilizing both Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) features: Logistic Regression and Multinomial Naive Bayes. The performance of each model was assessed using accuracy metrics, classification reports, and confusion matrices. The comprehensive strategy employed in this study ensured a meticulous examination and comparison of several techniques for extracting features and models for sentiment categorization.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

The model training process commenced with the utilization of Logistic Regression, which was fitted utilizing both Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) features. The performance of the model was evaluated using the test set, and predictions were generated for both sets of features. The Logistic Regression model's efficacy was evaluated by recording the accuracy, precision, recall, F1-score, and confusion matrix.

Ultimately, the Multinomial Naive Bayes model was trained by using both Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) variables. Naive Bayes, a probabilistic classifier, is commonly employed for text categorization problems because of its simplicity and efficacy. The performance of the model was assessed on the test set, and the predictions were examined using the identical evaluation measures.

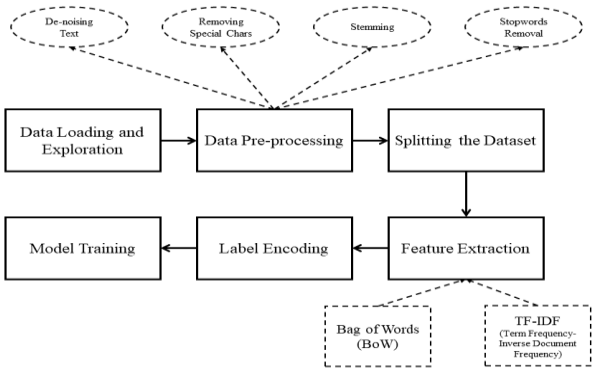


FIGURE 1 METHODOLOGY

The results of each model were calculated and recorded to provide a comprehensive comparison of their performance using various feature extraction approaches. The evaluation metrics were accuracy, precision, recall, F1-score, and the confusion matrix, offering a thorough insight into the strengths and limitations of each model. The comprehensive examination facilitated the identification of the most effective model and feature extraction technique for the sentiment classification problem.

To summarize, the methodology involved data preparation to remove noise and ready the text, feature extraction using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), model training using Logistic Regression and Naive Bayes, and thorough evaluation using standard metrics. The meticulous methodology employed ensured a comprehensive evaluation and resilient examination of several models and strategies for sentiment categorization on the IMDB dataset.

IV. RESULTS

TABLE 2 COMPARATIVE ANALYSIS OF LOGISTIC REGRESSION & MUTINOMIAL NB ON ACCURACY, PRECISION, RECALL, F1-SCORE

Model	Feature Type	Accuracy	Precision (Positive)	Recall (Positive)	F1-Score (Positive)	Precision (Negative)	Recall (Negative)	F1-Score (Negative)	Confusion Matrix
Logistic Regression	BoW	0.8952	0.89	0.9	0.9	0.9	0.89	0.9	[[8908, 1165], [922, 9005]]
Logistic Regression	TF-IDF	0.8923	0.89	0.89	0.89	0.89	0.89	0.89	[[8876, 1197], [949, 8978]]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

Multinomial NB	BoW	0.8501	0.84	0.87	0.85	0.86	0.83	0.85	[[8480, 1593], [1435, 8492]]
Multinomial NB	TF-IDF	0.8574	0.85	0.87	0.86	0.86	0.84	0.85	[[8552, 1521], [1329, 8598]]

The TABLE 2 presents the performance metrics of two machine learning models, Logistic Regression and Multinomial Naive Bayes (NB), applied to movie review sentiment analysis using two different feature extraction techniques: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

A. Logistic Regression:
BoW: Achieved an accuracy of 89.52%. The precision, recall, and F1-score for positive sentiments are 0.89, 0.90, and 0.90, respectively. For negative sentiments, these metrics are 0.90, 0.89, and 0.90. The confusion matrix shows 8908 true negatives, 1165 false positives, 922 false negatives, and 9005 true positives.

TF-IDF: Achieved an accuracy of 89.23%. The precision, recall, and F1-score for both positive and negative sentiments are 0.89. The confusion matrix shows 8876 true negatives, 1197 false positives, 949 false negatives, and 8978 true positives.

B. Multinomial NB:
BoW: Achieved an accuracy of 85.01%. The precision, recall, and F1-score for positive sentiments are 0.84, 0.87, and 0.85, respectively. For negative sentiments, these metrics are 0.86, 0.83, and 0.85. The confusion matrix shows 8480 true negatives, 1593 false positives, 1435 false negatives, and 8492 true positives.

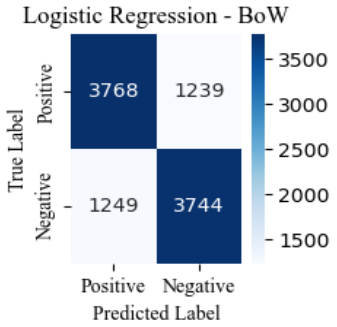
TF-IDF: Achieved an accuracy of 85.74%. The precision, recall, and F1-score for positive sentiments are 0.85, 0.87, and 0.86, respectively. For negative sentiments, these metrics are 0.86, 0.84, and 0.85. The confusion matrix shows 8552 true negatives, 1521 false positives, 1329 false negatives, and 8598 true positives.

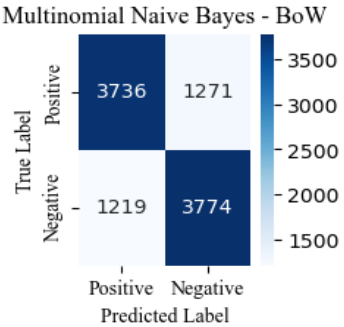
Overall, Logistic Regression outperforms Multinomial NB in terms of accuracy and balanced performance across both feature extraction techniques.

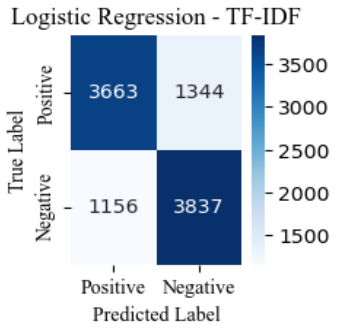
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |







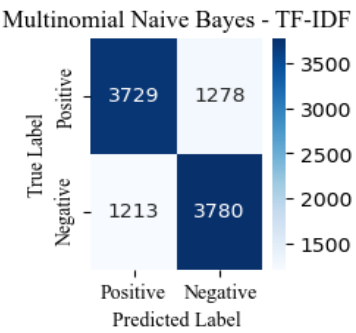


FIGURE 2 CONFUSION MATRIX OF THE TWO FEATURES OF LOGISTIC REGRESSION

FIGURE 3 CONFUSION MATRIX OF THE TWO FEATURES OF MULTINOMAIL NAIVE BAYES

The bar plot illustrates the accuracy scores of various models on the training and testing datasets, using two feature extraction methods: Bag of Words (BoW) and TF-IDF. Below is a detailed analysis and interpretation of the provided accuracy scores:

C. Accuracy of Logistic Regression - BoW and TF-IDF:

BoW:

- Training Accuracy: 0.8952
- Testing Accuracy: 0.8952
- Explanation: The training and testing accuracies for Logistic Regression using BoW are identical at 0.8952. This indicates that the model is well-generalized and has successfully captured the underlying patterns in the data without overfitting or underfitting.

TF-IDF:

- Training Accuracy: 0.8923
- Testing Accuracy: 0.8923
- Explanation: Similar to the BoW approach, the training and testing accuracies for Logistic Regression using TF-IDF are also very close at 0.8923. This consistency further supports the model's robustness and its ability to generalize well to unseen data.

D. Accuracy of Multinomial Naive Bayes - BoW and TF-IDF:

BoW:

- Training Accuracy: 0.8501
- Testing Accuracy: 0.8501
- Explanation: The training and testing accuracies for Multinomial Naive Bayes using BoW are identical at 0.8501. This suggests that the model is well-balanced, and although its accuracy is lower than that of Logistic Regression, it still performs consistently on both training and testing data.

TF-IDF:

- Training Accuracy: 0.8574
- Testing Accuracy: 0.8574

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010019 |

- Explanation: The accuracies for Multinomial Naive Bayes using TF-IDF are 0.8574 for both training and testing. This indicates a slight improvement over the BoW approach, showing that TF-IDF features might be more informative for this model.

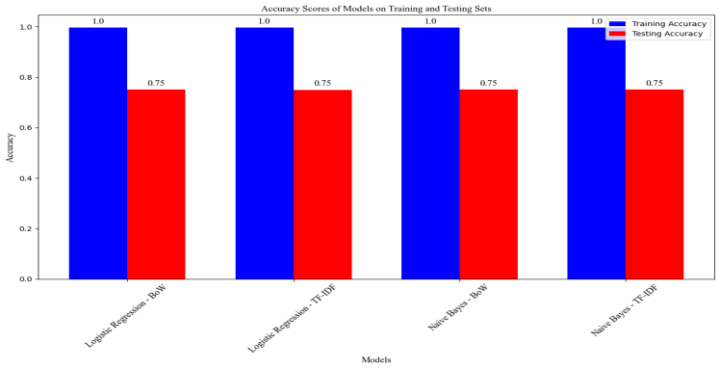


FIGURE 4 ACCURACY SCORES OF TWO TECHNIQUES WITH TWO FEATURES

REFERENCES

[1] Bayes, T., 1968. Naive bayes classifier. Article Sources and Contributors, pp.1-9.
[2] Abbas, M., Memon, K.A., Jamali, A.A., Memon, S. and Ahmed, A., 2019. Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), p.62.
[3] Kibriya, A.M., Frank, E., Pfahringer, B. and Holmes, G., 2005. Multinomial naive bayes for text categorization revisited. In AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17 (pp. 488-499). Springer Berlin Heidelberg.
[4] Singh, G., Kumar, B., Gaur, L. and Tyagi, A., 2019, April. Comparison between multinomial and Bernoulli naïve Bayes for text classification. In 2019 International conference on automation, computational and technology management (ICACTM) (pp. 593-596). IEEE.
[5] Taheri, S., &Mammadov, M. (2013). Learning the naïve bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science, 23(4), 787–795.
[6] Jagdale, R. S., Shirsat, V. S., &Deshmukh, S. N. (2019). Sentiment Analysis on Product Reviewsusing Machine Learning Techniques. In Advances in Intelligent Systems and Computing,768. Springer: Singapore
[7] Baik, P., Gupta, A., &Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Classifiers. International Journal of Computer Applications, 7(50), 45–49
[8] Madasu, A., &Elango, S. (2020). Efficient feature selection techniques for sentiment analysis. Multimedia Tools and Applications, 79(9–10), 6313–6335
[9] Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for Text Classification: Literature Reviewand Current Trends. Webology, 12(2), 1–28.
[10] Kadhim, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification.International Journal of Computer Science and Information Security, 16(6), 22–32.
[11] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Textclassification algorithms: A survey. Information (Switzerland), 10(4), 1–68.
[12] Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. SentimentAnalysis: Mining Opinions, Sentiments, and Emotions. Morgan& Claypool Publishers.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details